EXHIBIT 4-2

Am. J. Hum. Genet. 55:175–189, 1994

Evaluation of 13 Short Tandem Repeat Loci for Use in Personal Identification Applications

Holly A. Hammond,[1] Li Jin,[3] Y. Zhong,[3] C. Thomas Caskey,[1,2] and Ranajit Chakraborty[3]

[1]Department of Molecular and Human Genetics and [2]Howard Hughes Medical Institute, Baylor College of Medicine, and [3]Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston

Summary

Personal identification by using DNA typing methodologies has been an issue in the popular and scientific press for several years. We present a PCR-based DNA-typing method using 13 unlinked short tandem repeat (STR) loci. Validation of the loci and methodology has been performed to meet standards set by the forensic community and the accrediting organization for parentage testing. Extensive statistical analysis has addressed the issues surrounding the presentation of "match" statistics. We have found STR loci to provide a rapid, sensitive, and reliable method of DNA typing for parentage testing, forensic identification, and medical diagnostics. Valid statistical analysis is generally simpler than similar analysis of RFLP-VNTR results and provides powerful statistical evidence of the low frequency of random multilocus genotype matching.

Introduction

Identification of individuals is necessary in many social, medical, and forensic circumstances. The past 10 years of experience have demonstrated that DNA typing methodologies using a battery of highly polymorphic RFLP-VNTR probes (Jeffreys et al. 1985; Nakamura et al. 1987) and Southern-blot technology (Southern 1975) are far more efficient than the earlier protein typing methods, for resolving personal identification cases.

We have developed a PCR-based DNA typing method that uses common DNA polymorphisms to provide a rapid and reliable typing system for personal identification. Repeat sequences of two, three, four, and five bases are valuable polymorphic markers that can be used in identification and genetic linkage (Litt and Luty 1989; Tautz 1989; Weber and May 1989; Boylan et al. 1990; Weber 1990; Chakraborty and Kidd 1991; Edwards et al. 1991*a*, 1992; Huang et al. 1991; Caskey and Hammond 1992; Roewer and Epplen 1992). Since they are smaller than VNTR repeats, they can be informative when partially degraded DNA and/or limited DNA is available. These are frequently the limiting factors for VNTR analysis.

We present the development of a battery of short tandem repeat (STR) loci for use as a highly discriminatory system of genetic markers in personal identification, specifically for parentage testing, forensic identification, and medical applications. As validation of the utility of STR loci in these applications, we demonstrate conformation to Hardy-Weinberg expectations of genotype frequency data for samples obtained from four major population groups residing in Houston. We investigate the assumption of allelic independence across the loci, for validity in these populations, and present summary statistics indicating their potential for personal identification, along with indirect estimates of mutation rates for each locus.

Material and Methods

Population Groups

DNA was extracted (Patel et al. 1984) from blood samples collected from unrelated individuals presenting to a Houston area blood bank. Blood donors were visually designated as White, Black, or "other," by blood bank personnel. Mexican-American and Asian samples were identified from the "other" group by common ethnic first names and surnames. Ninety-seven to 386 chromosomes were examined for each of the four population groups.

Genetic Loci

The STR locus names, Genome Database (GDB) designations, chromosomal locations, PCR primers, and allele product sizes observed are given in table 1, for the 13 loci available in our personal identification system. We have designated the STR loci by their GenBank locus name and lowest alphabetical representation of the repeat region, following the lead of Edwards et al. (1991*a*), and/or by their GDB locus assignment. Alleles are designated by the number of full tandem repeats of the STR sequence. The number of repeats was determined by sequencing of PCR products for at least two alleles at each locus (authors' unpublished data).

Received September 21, 1993; accepted for publication March 11, 1994.

Address correspondence and reprints: Holly A. Hammond, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Room T809, Houston, TX 77030.

0002-9297/94/5501-0024$02.00

Table 1

STR Loci Studied

Genbank Locus[STR]$_n$/GDB Designation	Gene (chromosome location)	PCR Primers	Product Length (bp)
HUMHPRTB[AGAT]$_n$/HPRT	Hypoxanthine phosphoribosyltransferase (Xq26)	A: atg cca cag ata ata cac atc ccc B: ctc tcc aga ata gtt aga tgt agg	259–299
HUMFABP[AAT]$_n$/FABP2	Intestinal fatty acid–binding protein (4q28–q31)	A: gta gta tca gtt tca tag ggt cac c B: cag ttc gtt tcc att gtc tgt ccg	199–220
HUMCD4[AAAAG]$_n$/CD4	Recognition/surface antigen (cd4) (12p12-pter)	A: ttg gag tcg caa gct gaa cta gcg B: cca gga agt tga ggc tgc agt gaa	125–175
HUMCSF1PO[AGAT]$_n$/CSF1R	c-fms proto-oncogene for CSF-1 receptor (5q33.3-q34)	A: aac ctg agt ctg cca agg act agc B: ttc cac aca cca ctg gcc atc ttc	295–327
HUMTH01[AATG]$_n$/TH	Tyrosine hydroxylase (11p15.5)	A: gtg ggc tga aaa gct ccc gat tat B: att caa agg gta tct ggg ctc tgg	179–203
HUMPLA2A1[AAT]$_n$/PLA2A	Pancreatic phospholipase A-2 (12q23-qter)	A: ggt tgt aag ctc cat gag gtt aga B: ttg agc act tac tat gtg cca ggc t	118–139
HUMF13A01[AAAG]$_n$/F13A1	Coagulation factor XIII (6p24-p25)	A: gag gtt gca ctc gag cct ttg caa B: ttc ctg aat cat ccc aga gcc aca	281–331
HUMCYAR04[AAAT]$_n$/CYP19	Aromatase cytochrome P-450 (15q21.1)	A: ggt aag cag gta ctt agt tag cta c B: gtt aca gtg agc caa ggt cgt gag	173–201
HUMLIPOL[AAAT]$_n$/LPL	Lipoprotein lipase (8p22)	A: ctg acc aag gat agt ggg ata tag B: ggt aac tga gcg aga ctg tgt ct	125–175
HUMRENA4[ACAG]$_n$/REN	Renin (1q32)	A: aga gta cct tcc ctc ctc tac tca B: ctc tat gga gct ggt aga acc tga	255–275
HUMFESFPS[AAAT]$_n$/FES	c-fes/fps proto-oncogene (15q25-qter)	A: gct tgt taa ttc atg tag gga agg c B: gta gtc cca gct act tgg cta ctc	222–250
No GenBank entry/D6S366	Unknown (6q21-qter)	A: aga ggt tac agt gag ccg aga ttg B: gaa gtc cta aca gaa tgg aag gtc c	138–162
HUMARA[AGC]$_n$/AR	Androgen receptor (Xcen-q13)	A: tcc aga atc tgt tcc aga gcg tgc B: gct gtg aag gtt gct gtt cct cat	255–315

Genotype Determinations

Allele identification was achieved by PCR amplification of 10–50 ng of genomic DNA by using Perkin Elmer Cetus thermocyclers (models TC1 and 480), Amplitaq enzyme, and standard Cetus buffer conditions (10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin) in 15-μl volumes. The PCR products were radiolabeled by inclusion of 2 μCi [alpha-^{32}P]dCTP (3,000 Ci/mmol) in the PCR. Primer concentrations for all single-locus reactions and most arbitrary multiplexes were 1 μM for each primer. Standard sets of three loci were routinely analyzed in balanced triplex reactions. Concentrations for the balanced triplexes were determined experimentally. The goal was to achieve approximately equal radioactive signals on an autoradiograph, for all nine loci typed in three triplex PCR reactions. The final primer concentrations for these triplexes are as follows: multiplex 1—HUMHPRTB[AGAT]$_n$, 2.66 μM; HUMFABP[AAT]$_n$, 1.06 μM; and HUMCD4[AAAAG]$_n$, 0.67 μM; multiplex 2—HUMCSF1PO[AGAT]$_n$, 0.27 μM; HUMTH01[AATG]$_n$, 0.27 μM; and HUMPLA2A1-[AAT]$_n$, 1.40 μM; and multiplex 3—HUMF13A01-[AAAG], 0.27 μM; HUMCYAR04[AAAT]$_n$, 1.27 μM; and HUMLIPOL[AAAT]$_n$, 1.47 μM. The PCR conditions for all multiplexes and single-locus PCR reactions are as follows: 95°C, 2 min, 1 cycle; 95°C, 45 s, then 62°C, 30 s, then 72°C, 30 s, 28 cycles; and 72°C, 10 min, 1 cycle.

Allele separation of the radiolabeled PCR products was achieved by electrophoresis through a 4% (39:1) acrylamide-bisacrylamide denaturing (8 M urea) sequencing gel. The PCR products were diluted 1:1 with standard sequencing loading buffer. Two microliters of the diluted reactions were loaded onto a vertical PAGE gel system (IBI STS45 vertical gel apparatus) and electrophoresed at constant wattage (80 W) for 2–2½ h. After electrophoresis the gel plates were separated, and the gel was transferred to filter paper. The gel was exposed to Kodak XAR imaging film (with or without vacuum drying of the gel) for 4 h to overnight. Allele designations were made by comparison

to known alleles from two or more samples run in lanes adjacent to unknown samples.

Statistical Methods

Gene count estimates of allele frequencies, unbiased estimates of heterozygosity, and their standard errors (SEs) were calculated as described by Edwards et al. (1992). The expected number of heterozygotes for a locus-population combination was compared with the observed number by means of a χ^2 statistic to evaluate conformation to Hardy-Weinberg predictions. The significance was determined by 5,000 permutations of genotype shuffling to determine the proportion of times the resultant χ^2 exceeded the observed value. The expected number of distinct homozygous and heterozygous genotypes and their SEs were determined by previously published methods (Chakraborty 1993*a*, 1993*b*). Overall discordance of genotype frequencies from their Hardy-Weinberg expectations can be detected by these summary statistics. A likelihood-ratio (L) test criterion (*G*-statistic) was used to contrast observed and expected frequencies of all genotypes. This test criterion was computed by the methods of Sokal and Rohlf (1969) and Weir (1991). Its empirical significance was evaluated by a permutation method detailed by Chakraborty et al. (1991) and Deka et al. (1991). The exact (*E*) test (Guo and Thompson 1992) was also performed, to check the conformity of the observed genotype frequencies with their respective Hardy-Weinberg expectations.

Locus-specific fixation indices within individuals (i.e., inbreeding coefficients [ϕ]) for each population group and the pooled population were computed by the method of Yasuda (1968) using a maximum-likelihood procedure. The SE of these estimates, as well as tests of significant departure from $\phi = 0$ (Hardy-Weinberg equilibrium [HWE]), were conducted based on the entire multiethnic genotype data.

Pairwise independence of genotype frequencies for each combination of loci was tested by a procedure suggested by Risch and Devlin (1992) and Morton et al. (1993). The occurrence of matches at both loci, a match at one of the loci, or no match for two specific loci (A and B), was evaluated for all pairs of individuals, in combinations of two loci, in the observed multilocus data. Each pair of loci generated a 2×2 contingency table with observation n_{AB}, $n_{A\bar{B}}$, $n_{\bar{A}B}$, and $n_{\bar{A}\bar{B}}$, where n_{AB} represents the number of pairs of individuals whose genotypes for loci A and B are the same (match); $n_{A\bar{B}}$, represents the number of pairs that matched at locus A but not at locus B; $n_{\bar{A}B}$ represents the number of pairs that matched at locus B but not at locus A; and $n_{\bar{A}\bar{B}}$ represents the number of pairs that did not match at either locus. Traditional 2×2 contingency χ^2 analysis (Sokal and Rohlf 1969) indicates departure from independence. The significance of this χ^2 value was tested by genotype shuffling in the entire database. The empirical level of significance was judged by counting the proportion of times the χ^2 values in the shuffled data (2,000 replications) exceeded the observed χ^2. This test was performed for each pair of loci, resulting in 78 tests for each population group and the pooled data.

An approximate test of random association of alleles at different loci was performed as described by Edwards et al. (1992). The summary statistic s_k^2 (observed variance of the number of heterozygous loci) and its 95% confidence intervals were determined for the samples from each population group (Brown et al. 1980; Chakraborty 1984).

All of these tests presume that the individuals who exhibit only one allele product at a locus are truly homozygous. This may not necessarily be correct, since in PCR-based methods "null" alleles may not be detected, because of differential amplification or nonamplification caused by sequence polymorphisms within the primer sequence. While such possibilities have been empirically shown to occur at the dinucleotide repeat loci (Callen et al. 1993; Koorey et al. 1993), there is no direct molecular evidence that null alleles are indeed present at the STR loci studied here. The presence of null alleles will cause pseudodependence of alleles within, as well as across, loci in genotypes of individuals. Using modifications of a theory originally proposed by Gart and Nam (1984), we developed a method for estimating the null-allele frequency (r) based on the observed genotype (phenotype) frequency data (Chakraborty et al. 1994). Revised empirical levels of significance, incorporating the presence of null alleles, are calculated to examine whether either the significant deviations from HWE or the significant inbreeding coefficients could be attributed to null alleles alone.

The power of exclusion for each locus, as well as the combined powers of exclusion, were calculated as described by Chakravarti and Li (1983), for multiple-allele loci. In forensic science, a convenient measurement of the usefulness of a genetic marker is the individualization potential (P_I), which may be expressed as the sum of the squares of all genotype frequencies in a polymorphic system (Sensabaugh 1982). We present a comparison of the P_I values of all loci in four population groups, calculated under HWE and in the presence of inbreeding within a population. Indirect estimates of mutation rates for each of the 13 loci were obtained from the distribution of the number of segregating alleles at each locus in the populations sampled, using the maximum-likelihood method proposed by Chakraborty and Neel (1989).

Results

PCR Assay

Each locus described contains an STR. The polymorphisms are due to the variable numbers of repeats of three, four, and five nucleotides sequences (see table 1). We have found that three-, four-, and five-base repeats amplify more authentically and provide more easily interpretable

results than do dinucleotide repeat loci (Edwards et al. 1991*a*; Huang et al. 1991).

All repeat regions, with the exceptions of those within the HUMHPRTB, D6S366, and HUMCD4 loci, were identified by a computer search of GenBank sequences. The search criteria identified regions containing five or more repeats of each of the 44 unique three- and four-base sequences. The HUMHPRTB STR was identified during the sequencing of the hypoxanthine phosphoribosyltransferase gene in this laboratory (Edwards et al. 1990). D6S366 was identified by screening of a lambda bacteriophage library of human genomic DNA with oligonucleotides containing 10 repeats of the sequence AAT. Regions adjacent to the repeat were sequenced by previously published methods (Edwards et al. 1991*a*) and by direct sequencing of lambda bacteriophage clones (Panzer et al. 1993). The STR in the HUMCD4 locus was identified during sequencing of the CD4 gene (Edwards et al. 1991*b*).

Numerous loci identified as having five or more STR repeats were examined to find those that were highly polymorphic and that provided robust and easily interpreted signals (Edwards et al. 1991*a*). Some of these loci have been independently reported as polymorphic, while our population studies have been ongoing (Polymeropoulos et al. 1990, 1991*a*, 1991*b*, 1991*c*, 1991*d*, 1991*e*; Zuliani and Hobbs 1990; Edwards et al. 1991*b*; Hearne and Todd 1991; Ahn et al. 1992; Sleddens et al. 1992; Panzer et al. 1993). We have previously published database information and HWE determination for five of these loci (Edwards et al. 1992; Puers et al. 1993). A database of DNA samples from ~200 Caucasian, ~200 Black, ~200 Mexican-American, and ~80 Asian samples were typed for each locus. Amplifications and typing were performed in single and multiplex reactions. All loci amplified with 1 μM concentrations of the primers, at the conditions previously given.

Multiplex reactions were established to simplify analysis. Since primers were designed with ~24 bases and ~50% GC content, all amplification reactions have similar melting temperatures. Thus, all reactions can be performed under identical amplification conditions. Seven of the loci have been successfully amplified in one reaction. For utility, multiplexes were restricted to three loci. The amplification product size determined the groups of loci and their alleles that could be coamplified and resolved by sequencing gel separation.

Nine of the loci have been developed into three standard multiplex reactions of three loci each. The results of allele separation for these multiplexes were balanced with regard to the intensity of allele bands on the autoradiographs. Approximately equal autoradiography results were achieved by adjusting primer concentrations. Different primer concentrations are required for fluorescent or silver-staining detection. We report our conditions for autoradiography detection of deoxycytidine-5′-triphosphate, tetratriethylammonium salt, (alpha-^{32}P) incorporated into the PCR products.

Figure 1 shows a representative sample of the results for the three multiplex reactions and single-locus amplifications of alleles for the loci HUMFESFPS, and D6S366. Results for HUMARA and HUMRENA4 have been shown by Edwards et al. (1991*a*).

Validation and Nomenclature

All STR loci have been typed in DNA extracted from various tissues, body fluids, and cultured cells, including blood, cultured lymphoblasts, vaginal epithelial cells, semen, and skin tissue. Results of samples from the same individual gave identical allele typing for all tissues tested.

Inheritance patterns of the loci were tested on samples from families presenting to our diagnostic laboratory and from at least three three-generation CEPH families. All loci show codominant inheritance consistent with the expected pattern for the chromosome to which the locus has been assigned.

Allele designation follows the lead of Edwards et al. (1991*a*) in using the number of reiterations of the repeat sequences as the allele number. At least two alleles in each locus have been sequenced. Designations of the allele numbers assigned to the database and family studies were made by comparison to at least two external cell line controls, RJK 1094 and RJK 1258, electrophoresed simultaneously with samples of unknown allele designation. Part of the databases for the HUMF13A01 and HUMTH01 (Puers et al. 1993) loci were typed using an allele ladder provided by Promega Corporation. Several loci have shown variant alleles that do not correspond to a change in the number of reiterations of the core sequence. Where the sequence of these alleles is known, they have been designated per the recommendation on nomenclature standardization of STR loci, made by the DNA commission members at the meeting of the International Society of Forensic Haemogenetics, in Venice, Italy, in October 1993. In the two cases in which the sequence difference of the variant allele has not been determined (HUMHPRTB [Edwards et al. 1992] and HUMCYAR04), the variant allele has been designated as "*N*A," where *N* is the allele number of the closest, smaller allele.

Allele Frequency Estimates

A complete enumeration of all genotype frequencies for such hypervariable loci requires substantial space, and such data are not easily comprehensible without appropriate summarization. As an illustration, table 2 shows the genotype frequencies, for the Caucasian data, at the D6S366 locus. This locus exhibits a moderate heterozygosity (70%–80%) and has a large number of genotypes. This is typical of all of the loci. The gene count estimates of allele frequencies for eight of the loci are given in Ap-

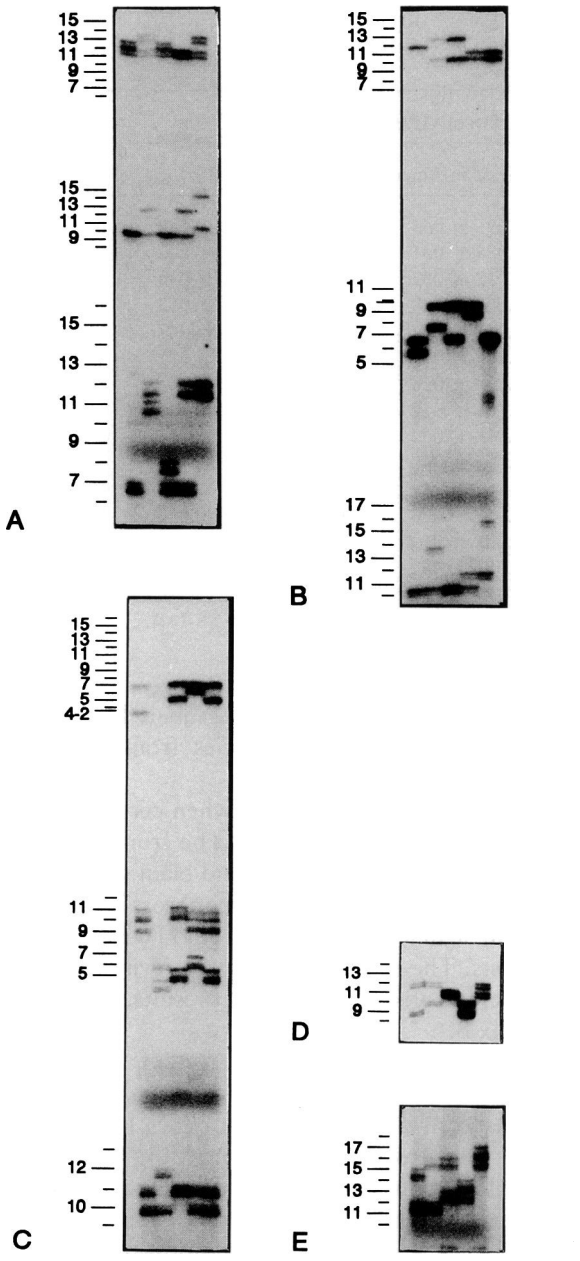

Figure 1 Examples of autoradiographs showing separation of PCR products from three three-locus multiplexes and two single-locus PCR reactions, on a denaturing sequencing gel. Representative alleles are shown for (*A*) multiplex 1—HUMHPRTB[AGAT]$_n$ (*top*), HUMFABP[AAT]$_n$ (*middle*), and HUMCD4[AAAAG]$_n$ (*bottom*); (*B*) multiplex 2—HUMCSF1PO[AGAT]$_n$ (*top*), HUMTH01[AATG]$_n$ (*middle*), and HUMPLA2A1[AAT]$_n$ (*bottom*); (*C*) multiplex 3—HUMF13A01[AAAG]$_n$ (*top*), HUMCYAR04[AAAT]$_n$ (*middle*), and HUMLIPOL[AAAT]$_n$ (*bottom*); (*D*) HUMFESFPS[AAAT]$_n$; and (*E*) D6S366. The first two lanes of each set of reactions contain the controls RJK 1094 and RJK 1258. The allele designations are shown as a scale next to each set of alleles. The nonuniform tick marks on the scales in panel *B* (*middle*) and panel *C* (*top*) designate the variant alleles that are shown in the photograph.

Table 2

Genotype Frequencies of D6S366 Caucasians

	Allele Number								
	10	11	12	13	14	15	16	17	18
10	...	...	...	1	1	2	...	1	...
11		1	2	3	5	2	4	1	...
12			20	12	7	10	15	4	...
13				8	10	9	13	3	1
14					1	6	6	1	...
15						2	10	...	...
16							6	3	...
17								...	...
18									...

Note.—Total no. of individuals = 170.

pendix A. Similar data on the other five loci have been published previously (Edwards et al. 1992; Puers et al. 1993).

Tests of HWE

HWE depends on the existence of large randomly mating (panmictic) populations. These characteristics are not always found in natural populations. Deviations from the predictions of the square law, $(\Sigma p_i)^2 = 1$, may be caused by a variety of factors including nonrandom mating, ascertainment bias, distortions in the contribution of gametes to offspring, and faulty data interpretation (Milkman and Beatty 1970; Bell et al. 1982). The calculations of genotype frequency from allele frequencies require approximate agreement with the square law.

Three tests were performed to meet this requirement: the χ^2 test based on overall heterozygosity, the L test, and the E test. These tests were designed to indicate deviation of the observed genotype frequencies from those expected under HWE. The observed and expected number of heterozygotes, on which the χ^2 test is constructed, and the observed and expected number of distinct genotypes (data not shown) provide summary statistics that detect gross deviation from equilibrium. The L test (Weir 1991) was used to contrast the observed and expected frequencies of all possible genotypes. Since the values of these test statistics (χ^2 or L) are, by themselves, not indicative of departure from HWE, and since the E-test procedure of Guo and Thompson (1992) does not provide a comparable test statistic, the empirical levels of significance for each test and locus-population combination are contained in Appendix B. The levels of significance for χ^2 and L were obtained by the permutation method as described by Chakraborty et al. (1991). Significant deviation from the predictions of Hardy-Weinberg for each test-locus-population combination are indicated in Appendix B.

Deviations were detected in 8 of the 52 possible locus-population combinations, by comparison of the observed

Table 3

Estimates of Fixation Indices within Individuals, and their SEs

Locus	Population				
	Caucasian	Black	Mexican-American	Asian	Pooled
HUMHPRTB	.116 ± .046[a]	.062 ± .040	.096 ± .065	.092 ± .087	.056 ± .021[a]
HUMFABP	−.003 ± .031	−.037 ± .027	.038 ± .031	.088 ± .047	.032 ± .015[a]
HUMCD4	−.005 ± .032	−.003 ± .024	.012 ± .032	.149 ± .043[a]	.026 ± .012[a]
HUMCSF1PO	−.001 ± .031	−.025 ± .026	−.038 ± .028	−.024 ± .053	−.010 ± .014
HUMTH01	.059 ± .030[a]	−.047 ± .033	.051 ± .032	.063 ± .047	.058 ± .015[a]
HUMPLA2A1	−.030 ± .030	−.021 ± .028	−.005 ± .028	.173 ± .051[a]	.027 ± .015
HUMF13A01	.033 ± .027	.016 ± .022	−.024 ± .028	.010 ± .051	.035 ± .011[a]
HUMCYAR04	.036 ± .027	.033 ± .029	−.006 ± .028	−.047 ± .047	.032 ± .014[a]
HUMLIPOL	−.044 ± .036	.003 ± .031	−.018 ± .033	−.049 ± .051	−.002 ± .016
HUMRENA4	−.010 ± .043	.014 ± .033	−.005 ± .042	.039 ± .067	.017 ± .018
HUMFESFPS	−.021 ± .033	.058 ± .032	.017 ± .032	.077 ± .050	.046 ± .016[a]
D6S366	.015 ± .027	.018 ± .026	−.016 ± .026	.063 ± .050	.026 ± .014
HUMARA	.012 ± .032	.067 ± .027[a]	−.008 ± .040	.121 ± .050[a]	.048 ± .015[a]
Weighted average	.011 ± .032	.008 ± .028	−.001 ± .032	.060 ± .051	.029 ± .014

[a] Signifiant deviation from $\phi = 0$.

and expected number of heterozygotes (χ^2). The deviations occurred in the HUMHPRTB[AGAT]$_n$-Caucasian, HUMHPRTB[AGAT]$_n$-Black, HUMTH01[AATG]$_n$-Asian, HUMPLA2A1[AAT]$_n$-Asian, HUMF13A01[AAAG]$_n$-Caucasian, HUMCYAR04[AAAT]$_n$-Black, HUMFESFPS[AAAT]$_n$-Asian, and HUMARA[AGC]$_n$-Black populations. The L test showed only three deviations, at HUMTH01[AATG]$_n$-Black, HUMF13A01[AAAG]$_n$-Black, and HUMCYAR04[AAAT]$_n$-Black, in 52 possible combinations. The L-test statistic was also barely insignificant (P = 5.2%, with 5,000 replications of permutations) for the D6S366-Black combination. The E test revealed significant departure from Hardy-Weinberg proportions in 7 of the 52 combinations examined: HUMTH01[AATG]$_n$-Black, HUMTH01[AATG]$_n$-Asian, HUMARA[AGC]$_n$-Black, HUMPLA2A1[AAT]$_n$-Asian, HUMFESFPS[AAAT]$_n$-Asian, D6S366-Black, and HUMCYAR04[AAAT]$_n$-Black. In addition, comparisons of the observed and expected number of distinct genotypes (data not shown) detected only one deviation (HUMCD4[AAAAG]$_n$-Asian). When each of these tests was performed on the pooled sample, significant deviations were noted by one or more of the tests, for 10 of the 13 loci (last column of Appendix B).

It should be noted that the deviating locus-population combinations were not the same for each of the three tests. Deviation from the expectations of Hardy-Weinberg was observed in most of the pooled population where deviation would be expected. Only two of the locus-population combinations (HUMARA[AGC]$_n$-Black and HUMCYAR04[AAAT]$_n$-Black) demonstrated consistent deviation from the expectations of Hardy-Weinberg proportions. These data indicate, that with the possible exception of the HUMARA[AGC]$_n$-Black and HUMCYAR04-[AAAT]$_n$-Black combinations, the square law may be used to calculate genotype frequencies from the allele frequencies.

These conclusions hold true when certain fractions of the homozygotes are assumed to be from nondetectable null-allele heterozygosity. Gart and Nam's (1984) estimate of r, incorporated in allele shuffling across individuals, produced no significant deviation from HWE, by any of the three tests. The estimated r and the revised levels of significance, for originally significant results, are indicated in Appendix B.

ϕ and Population Substructure

Estimates of ϕ and their SEs for each locus-population combination are presented in table 3. Examination of these values allows evaluation of whether the significant departure from HWE (Appendix B) is due to population substructure (Yasuda 1968). Footnotes to the table indicate significant departure of ϕ from HWE. The last row of the table contains the weighted averages of values across loci, for each population. Of the 52 estimates in the four samples, 6 are significantly different from zero (2 in Caucasians, 1 in Blacks, and 3 in Asians). There is no correlation between locus-population samples showing deviation from HWE in Appendix B and the data in table 3. Eight of the 13 loci show significant fixation indices in the pooled sample. Although the average degree of inbreeding (across loci) is not significant in any of the ethnic samples, the Asian sample shows the largest average fixation index (ϕ = 5.96%). Average fixation indices in the Caucasian, Black, and Hispanic populations are at the level of $\leq$1%, consistent with the observation in VNTR loci (Morton et al. 1993). The estimate of ϕ (3%) in the pooled sample shows

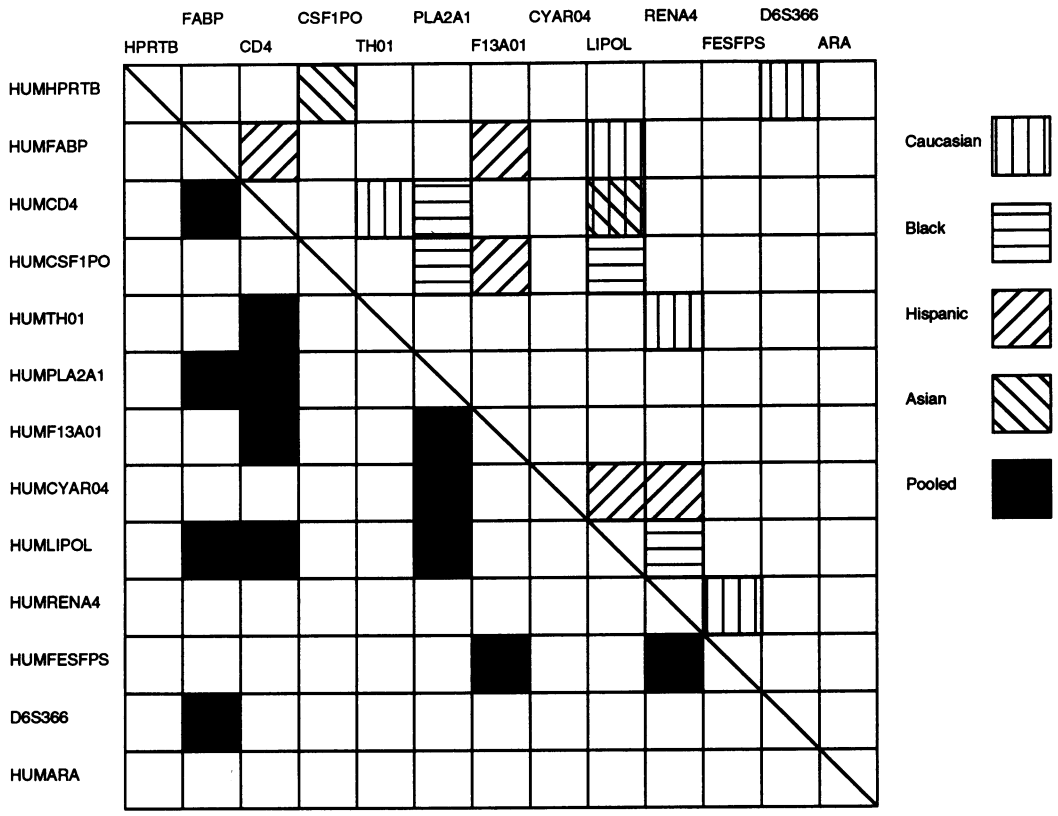

Figure 2 Illustration of the significant dependence of alleles in pairs of loci. Hatches in the upper diagonal indicate the populations that show dependence of alleles for each pair of loci. Shading in the lower diagonal indicates the pooled populations that showed dependence of alleles between pairs of loci.

that the effect of substructuring on the genotype probabilities is not appreciable.

Since the presence of null alleles may also be responsible for nonsignificant ϕ estimates, we calculated the correlation coefficients of the locus-specific ϕ's and r's within each population $\rho = .95, .92, .96$, and $.91$, for Caucasians, Blacks, Mexican-Americans and Asians, respectively). These suggest that the presence of related individuals in the sample, or hidden substructure within the populations sampled, may not be the only causes of the individual significant locus-population ϕ's. We show that the effects of the observed ϕ values on using the allele frequency data for personal identification purposes is not appreciable.

Independence across Loci

Substructuring within populations can also produce gametic phase disequilibrium. The random association of alleles between genetically unlinked loci is disturbed by population substructure. Figure 2 shows the pairs of loci where genotype probabilities are not independent, as determined by the contingency-table χ^2 test (empirical) based on two-locus match frequency observations. In the upper-diagonal segment, we have 17 significant pairs of loci (6 in Caucasians, 4 in Blacks, 5 in Mexican-Americans, and 2 in Asians), which is close to the expectation (15.6), based on 5% level of significance, for 312 total tests. Thirteen of the 78 pairwise tests showed significant evidence of dependence of pairwise genotype frequencies in the pooled samples (lower-diagonal segment of fig. 2).

Table 4 presents the values of the s_k^2 statistic for the loci studied. The values do not differ significantly from the expected values. The failure of the statistic to detect amalgamation in the pooled population indicates an absence of linkage disequilibrium, or lack of power in the test, as discussed by Edwards et al. (1992).

We observed nearly the expected number of significant pairwise χ^2 values, and no deviation from linkage equilibrium based on the s_k^2 statistic. These observations suggest that even if substructure exists within any of the four sampled populations, its level is so low that it has virtually no effect on the multilocus genotype probabilities.

Mutation Rates

Indirect estimates of mutation rates for the 13 STR loci are shown in table 5. The mutation rate estimates were between 2.36×10^{-5} and 1.86×10^{-4} for the STR loci. This compares favorably with previously reported rates

Table 4

Test for Random Association of Alleles[a] at 13 STR Loci in Four Populations

Population	Eleven Autosomal Loci				All 13 Loci[b]			
	n[c]	Observed	Expected	95% CI[d]	*n*	Observed	Expected	95% CI[d]
Caucasian	127	2.39	2.20	1.68–2.73	51	2.47	2.48	1.53–3.42
Black	110	1.49	1.94	1.44–2.45	50	1.72	2.32	1.43–3.22
Mexican-American	104	2.22	2.21	1.63–2.78	28	2.62	2.44	1.19–3.70
Asian	45	2.92	2.46	1.48–3.43	16	2.73	2.72	1.89–4.56
Pooled	386	2.36	2.22	1.91–2.52	145	2.42	2.52	1.95–3.08

[a] The statistic s_t^2 (the observed variance of the no. of heterozygous loci per population) detects linkage disequilibrium of alleles from the distribution of heterozygous alleles in a sample (Sokal and Rohlf 1969).
[b] For female samples, the test was performed only in the X-linked loci.
[c] No. of individuals studied.
[d] 95% confidence intervals.

(Edwards et al. 1992) and is generally between the values estimated for protein markers and VNTR loci.

Power of Exclusion

Parentage studies report two statistical values: the power of exclusion for the loci studied and the probability of paternity, based on the alleles shared between parent and child at each locus. The average power of exclusion was calculated for each locus, for all four population groups. The powers of exclusion ranged from 19.01% to 80.72%. The calculated values for the Caucasian population are presented in table 5. Combined average powers of exclusion for each population are depicted in figure 3. The order of typing is the listed order of loci in table 1.

Calculations for Forensic Evaluations

Forensic identifications usually require comparison of the typing results from a known sample to those of a sample of unknown origin. A statement can then be made as to the possibility that the unknown sample could have originated from the individual donating the known sample. This comparison is straightforward, and DNA typing of VNTR loci and our STR loci provide very high power of discrimination between unrelated individuals. When an unknown sample is declared to match in genotype for all typed loci, the forensic scientist is then asked to place a significance number on this "match." Typically the forensic scientist will report the frequency at which the com-

Table 5

Statistics for Forensic Identification and Parentage Studies

Locus	Het[a] (%)	P_{ex}[b]	$\nu \pm SE$[c]
HUMHPRTB[AGAT]$_n$	77.2	56.0	$4.84 \times 10^{-5} \pm 9.86 \times 10^{-6}$
HUMFABP[AAT]$_n$	64.9	40.7	$4.63 \times 10^{-5} \pm 8.46 \times 10^{-6}$
HUMCD4[AAAAG]$_n$	68.0	39.8	$4.81 \times 10^{-5} \pm 8.60 \times 10^{-6}$
HUMCSF1PO[AGAT]$_n$	74.2	50.3	$4.88 \times 10^{-5} \pm 8.75 \times 10^{-6}$
HUMTH01[AATG]$_n$	77.3	56.0	$3.96 \times 10^{-5} \pm 7.66 \times 10^{-6}$
HUMPLA2A1[AAT]$_n$	72.4	51.4	$4.61 \times 10^{-5} \pm 8.40 \times 10^{-6}$
HUMF13A01[AAAG]$_n$	73.3	49.4	$6.60 \times 10^{-5} \pm 1.06 \times 10^{-5}$
HUMCYAR04[AAAT]$_n$	72.7	48.9	$4.64 \times 10^{-5} \pm 8.46 \times 10^{-6}$
HUMLIPOL[AAAT]$_n$	68.1	41.7	$3.57 \times 10^{-5} \pm 7.21 \times 10^{-6}$
HUMRENA4[ACAG]$_n$	36.4	19.0	$2.36 \times 10^{-5} \pm 5.62 \times 10^{-6}$
HUMFESFPS[AAAT]$_n$	70.2	44.3	$4.30 \times 10^{-5} \pm 8.12 \times 10^{-6}$
D6S366	82.5	64.8	$6.13 \times 10^{-5} \pm 1.02 \times 10^{-5}$
HUMARA[AGC]$_n$	88.6	76.3	$1.86 \times 10^{-4} \pm 2.53 \times 10^{-5}$

[a] Unbiased estimate of heterozygosity of the Caucasian population.
[b] Power of exclusion for the Caucasian population.
[c] Indirect estimate of mutation rate and SE for each STR locus (Chakraborty and Neel 1989).

Table 6

Individualization Potential Values Calculated under HWE and in the Presence of Inbreeding (ϕ)

	Population									
	Caucasian		Black		Mexican-American		Asian		Pooled	
	HWE	ϕ	HWE	ϕ	HWE	ϕ	HWE	ϕ	HWE	ϕ
HUMHPRTB[AGAT]$_n$	.0853	.0816	.0865	.0836	.1252	.1221	.1580	.1615	.0917	.0893
HUMFABP[AAT]$_n$	.1715	.1715	.0657	.0672	.2400	.2397	.2315	.2316	.1258	.1262
HUMCD4[AAAAG]$_n$	.1693	.1699	.0532	.0533	.1700	.1688	.2842	.2731	.1103	.1091
HUMCSF1PO[AGAT]$_n$	.1129	.1130	.0817	.0831	.1249	.1286	.1112	.1131	.1045	.1052
HUMTH01[AATG]$_n$	.0866	.0843	.0950	.0978	.0958	.0931	.1201	.1190	.0759	.0732
HUMPLA2A1[AAT]$_n$	.1083	.1085	.0558	.0567	.0823	.0825	.0732	.0697	.0645	.0636
HUMF13A01[AAAG]$_n$	.1182	.1162	.0647	.0643	.0765	.0780	.1774	.1770	.0709	.0693
HUMCYAR04[AAAT]$_n$	.1214	.1193	.1602	.1580	.1473	.1478	.1301	.1336	.1213	.1190
HUMLIPOL[AAAT]$_n$	.1608	.1652	.1011	.1009	.1923	.1931	.2757	.2724	.1477	.1478
HUMRENA4[ACAG]$_n$	.4341	.4332	.3431	.3445	.4108	.4104	.4057	.4096	.3901	.3919
HUMFESFPS[AAAT]$_n$	.1448	.1468	.1013	.0986	.1445	.1433	.1429	.1408	.1230	.1206
HUMARA[AGC]$_n$	.0253	.0251	.0168	.0165	.0220	.0221	.0265	.0264	.0159	.0156

bined genotype will be observed in a population (i.e., 1 in 100 people, 1 in 1 million people, etc.). This is done by calculating the individual genotypes and multiplying genotype frequencies across unlinked loci. Figure 4 shows the cumulative genotype frequency of the most common genotype for each locus analyzed, in four population groups. Loci were accumulated in the order in which they are listed in table 1. Seven STR loci provide a most common genotype frequency of <1 in 10,000 people, for all four population groups.

Table 6 compares the values of the individualization potential calculated under HWE and in the presence of inbreeding within populations. The estimated values of ϕ from table 3 were used in the calculations. The difference in the P_I values is <5%. The presence of inbreeding within the population does not contribute significantly to the calculations of either multilocus statistics or genotype frequencies.

Discussion

We have established a DNA-typing methodology using 13 STR loci and have performed a genotype survey of four

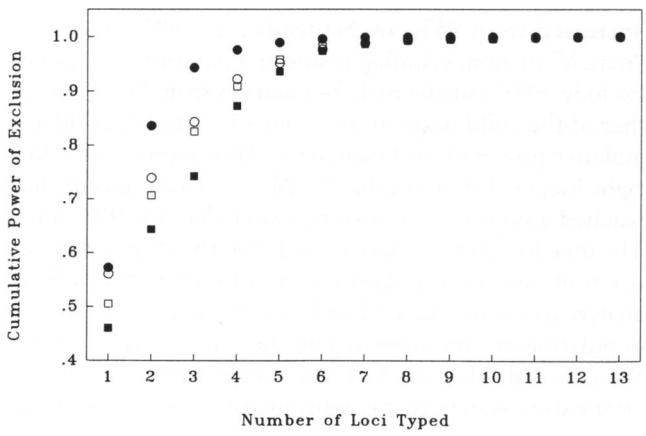

Figure 3 Graph of the cumulative power of exclusion for four populations, by number of loci tested. ○ = Caucasian power-of-exclusion data; ● = Black data; □ = Mexican-American data; and ■ = Asian data. All cumulative powers of exclusion are >99% after typing of the first eight STR loci.

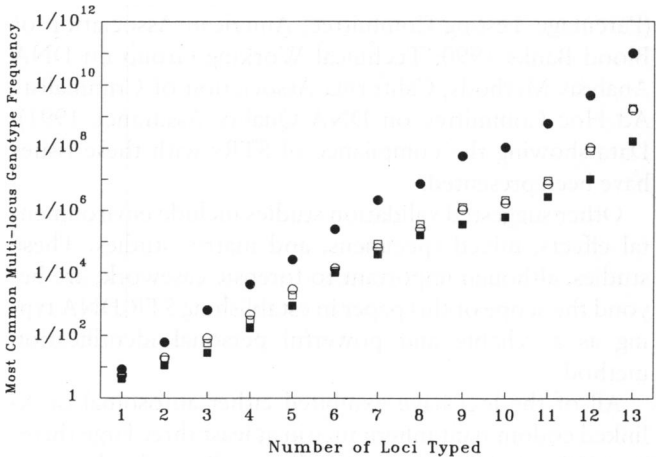

Figure 4 Graph of the most common multilocus genotype frequency for four populations, by number of loci tested. ○ = Caucasian multilocis-genotype data; ● = Black data; □ = Mexican-American data; and ■ = Asian data. A frequency of 1 in 1 million will be obtained for most individuals after typing the first nine STR loci.

population groups, using these loci. The validity of using the square law (HWE) to calculate expected genotype frequencies and of employing the product rule to combine genotype frequencies across loci has been examined.

The populations studied were selected on the basis of both visual identification of population group and surnames. While all individuals may not be accurately classified, improper classifications are likely to be small in number. This classification system also reflects the classification of individuals as they might be described for a forensic or parentage-testing case. We have only classified the individuals into broad racial/ethnic population groups.

We have presented information validating the STR typing method for both parentage testing and forensic use. Criteria suggested for choosing loci for DNA typing have been suggested (Parentage Testing Committee, American Association of Blood Banks 1990; Technical Working Group on DNA Analysis Methods, California Association of Criminalists Ad Hoc Committee on DNA Quality Assurance 1991). The main questions for validating STR typing for personal identification are as follows: Do the loci exhibit Mendelian inheritance consistent with the chromosome on which they lie? Is the mutation rate sufficiently low (<0.2%; Parentage Testing Committee, American Association of Blood Banks 1990) that new mutations will be rare? Do the populations studied provide allele frequencies consistent with the expectations of Hardy-Weinberg, so that they may be used to calculate genotype frequencies and probabilities of paternity? Are the loci studied unlinked, so that the product rule can be used to accumulate statistical information across loci? Are the powers of exclusion large enough to provide significant parentage-testing results after typing a reasonable number of loci? Are STR DNA typing results the same for standard specimens of varying age? Does DNA isolated from various tissues of the same individual yield the same genotyping result? (Parentage Testing Committee, American Association of Blood Banks 1990; Technical Working Group on DNA Analysis Methods, California Association of Criminalists Ad Hoc Committee on DNA Quality Assurance 1991). Data showing the compliance of STRs with these issues have been presented.

Other suggested validation studies include environmental effects, mixed specimens, and matrix studies. These studies, although important to forensic casework, are beyond the scope of this paper in establishing STR DNA typing as a reliable and powerful personal identification method.

All of the loci have exhibited either autosomal or X-linked codominant inheritance in at least three large three-generation families. Other, smaller family studies have not shown variance from this observation.

The population studies and tests for HWE, although showing some deviation for particular locus-population-test combinations, do not show a consistent pattern of deviation. Only 18 deviations were seen in the 156 possible locus-population-test combinations presented in this paper. Previous presentation of data for five of the loci showed similar results (Edwards et al. 1992). The few differences between the data presented in this paper for the five STR loci previously examined (Edwards et al. 1992) have three bases. In Edwards et al. (1992) the HUMHPRTB[AGAT]$_n$ and HUMFABP[AAT]$_n$ loci contained data from the CEPH families. The data presented here do not include the CEPH samples. A new permutation analysis has been performed (each new analysis with random permutations will produce slightly different probabilities for each statistical test), and 5,000 permutations (as compared with 1,000 in Edwards et al. 1992) were performed. Since no consistency in the deviations was observed, we must conclude that the loci presented are in HWE.

We have further shown that, between loci, there is no random association of the alleles. They are in linkage equilibrium. Pairwise comparison of multigenotype data does not show consistent significant dependence of alleles, between any of the loci. These results are not surprising, since the majority of the loci reside on separate chromosomes. Thus, the product rule for combining genotype frequencies across unlinked loci is valid for use with the loci we have studied.

Additional study of the effect of inbreeding within populations suggests that substructuring is not present at detectable levels, since, on average, none of the populations exhibits a significant ϕ. If substructuring exists in the populations, its effect is too small to contribute significantly to multilocus genotype frequencies. This would indicate that the four major population groups are appropriate classifications. Allele frequency data for each locus-population will provide reasonable estimates of multilocus genotype frequencies by product rule.

In Texas, the family-law courts require any method of paternity testing to provide a power of exclusion of ≥99% (increased from 95%, on September 1, 1993). That is, a "match" or nonexcluding result in a paternity case, must exclude ≥99% of the male population from being the father of the child in question. Figure 3 summarizes the cumulative power of exclusion data. After typing of the first eight loci (as listed in table 1), all population groups had reached a power of exclusion greater than the 99% limit. The nine loci that we have included in the three multiplex reactions are our standard battery of loci with which we analyze paternity cases. The power of exclusion, in all four population groups, after typing these nine loci, is >99%. We conclude that the STR loci do provide a highly discriminatory system that is appropriate for use in parentage testing.

In addition to the statistical analyses presented, all 13 loci were examined with respect to the mode of allele generation. With exclusion of the Asian data (insufficient sample size), we found 2 of 39 possible deviations from con-

formity, with the stepwise mutation model. Thirteen of the 39 locus-population combinations showed deviation from conformity to the infinite-allele model (IAM) (Shriver et al. 1993). As previously observed for the five loci studied by Edwards et al. (1992), STR loci (3-, 4-, and 5-nucleotide repeats) more closely emulate the stepwise mutation model than the IAM. The indirect estimates of the mutation rates presented here (see table 5), which assume the IAM, are overestimates even if the STR loci strictly follow the stepwise-mutation model. Since the largest of these estimates (1.86×10^{-4} for the HUMARA[AGC]$_n$ locus) is an order of magnitude smaller than that for the D1S7 minisatellite locus (Jeffreys et al. 1988), we conclude that the estimates of mutation rates for the 13 loci included in the present study are sufficient to indicate that an isolated discordance of genotypes, between a putative parent-offspring relationship, is the result of a de novo mutation.

Calculations of fixation indices within individuals (i.e., ϕ) for each locus-population combination have shown that the levels of departure from HWE cannot be explained sufficiently by the presence of substructure within the populations. Further, using ϕ estimates does not significantly change calculations of the individualization potential or "match probability." Estimates of genotype frequencies will not be significantly changed by including ϕ values in their calculation. The frequency of the homozygous genotype with the most common allele in all 52 locus-population combinations (HUMRENA4[ACAG]$_n$, Caucasian allele 8, 78.3 ± 2.2% [Edwards et al. 1992]) shows a <1% underestimate of the frequency when only HWE, rather than HWE with ϕ values, is used. This difference will have very little impact on a multilocus genotype having a frequency $\leq 1.0 \times 10^{-6}$ (1 in 1 million). Similar calculations with the power-of-exclusion values and probabilities of paternity show equally small differences in the multilocus values.

The benefit of STR typing over existing VNTR loci that have similar characteristics is the ease and speed of analysis. Nine to 13 (or more) STR loci can be typed in 2 d, from prepared DNA. The use of PCR for amplification of the loci also decreases the amount of DNA that must be used for testing. Degraded DNA samples are also amenable to the STR methods. Many samples that would fail to provide VNTR typing results will produce STR results. This system is also amenable to automation using fluorescent detection techniques (Edwards et al. 1991*a*, 1992). We have successfully used this technology to resolve questions involving laboratory sample switches, contamination of cultured amniocytes or chorionic villi cells in prenatal diagnosis, questioned parentage, bone marrow transplants, missing children, aircraft accidents, sexual assault, murder, other criminal identification issues, and identification of war casualties. In conclusion, we find that the trimeric, tetrameric, and pentameric STR loci we have presented provide an accurate, highly discriminating, sensitive, and rapid technique for DNA typing for personal identification issues in forensic science, parentage testing, and medical diagnostics.

Acknowledgments

We thank Dr. Belinda Rossiter for reviewing the manuscript. C.T.C. is an investigator with the Howard Hughes Medical Institute. This work was prepared under grants 90-IJ-CX-0037 and 92-IJ-CX-K042 (to C.T.C.) and 90-IJ-CX-K024 (to R.C.), from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view or opinions in the document are those of the authors and do not necessarily represent the official position of the Department of Justice.

Appendix A

Table A1

Allele Frequencies for Eight STR Loci in Four Populations

	Allele Number, Frequency ± Standard Error at Locus							
	HUMCD4	HUMCSF1PO	HUMPLA2A1	HUMF13A01	HUMCYAR04	HUMLIPOL	HUMFESFPS	D6S366
Caucasians:								
	7, 38.7 ± 2.5	8, .3 ± .3	10, 1.9 ± .7	3.2,[a] 8.3 ± 1.5	5, 34.2 ± 2.4	9, 3.7 ± 1.0	8, 1.4 ± .6	10, 1.5 ± .6
	8, 30.6 ± 2.4	9, 4.0 ± 1.1	11, 46.2 ± 2.6	4, 2.0 ± .8	6, 15.0 ± 1.8	10, 43.6 ± 2.6	10, 30.7 ± 2.4	11, 5.6 ± 1.2
	11, .3 ± .3	10, 26.4 ± 2.4	12, 13.0 ± 1.7	5, 19.2 ± 2.1	7, 10.0 ± 1.5	11, 28.3 ± 2.3	11, 39.0 ± 2.6	12, 26.5 ± 2.4
	12, 27.7 ± 2.3	11, 26.7 ± 2.4	13, 1.9 ± .7	6, 34.5 ± 2.5	8, .5 ± .4	12, 22.0 ± 2.1	12, 22.0 ± 2.2	13, 20.0 ± 2.2
	13, 2.1 ± .7	12, 33.4 ± 2.5	14, 13.2 ± 1.8	7, 32.5 ± 2.5	9, 1.3 ± .6	13, 2.4 ± .8	13, 6.3 ± 1.3	14, 11.2 ± 1.7
	14, .5 ± 0.4	13, 7.0 ± 1.4	15, 14.6 ± 1.8	8, .6 ± .4	10, 35.3 ± 2.4		14, .5 ± .4	15, 12.6 ± 1.8
		14, 2.0 ± .8	16, 9.1 ± 1.5	11, .3 ± .3	11, 2.6 ± .8			16, 18.5 ± 2.1
		15, .6 ± .4		14, .9 ± .5	12, 1.0 ± .5			17, 3.8 ± 1.0
				15, 1.7 ± .7				18, .3 ± .3
	n 382	344	370	348	380	378	364	340
Blacks:								
	7, 31.1 ± 12.4	7, 5.2 ± 1.2	10, .5 ± .4	3.2,[a] 11.1 ± 1.7	5, 36.6 ± 2.6	7, .6 ± .4	7, .3 ± .3	9, 3.1 ± .9
	8, 13.2 ± 1.7	8, 4.6 ± 1.1	11, 10.6 ± 1.6	4, 5.7 ± 1.2	6, 42.5 ± 2.6	9, 14.1 ± 1.9	8, 7.3 ± 1.4	10, 1.4 ± .6
	9, .8 ± .4	9, 2.7 ± .8	12, 13.2 ± 1.8	5, 34.0 ± 2.5	6A,[a] 1.4 ± .6	10, 35.9 ± 2.6	9, 4.6 ± 1.2	11, 6.7 ± 1.3
	10, 14.5 ± 1.8	10, 25.5 ± 2.3	13, 12.8 ± 1.7	6, 14.6 ± 1.9	7, 5.5 ± 1.2	11, 13.5 ± 1.8	10, 24.0 ± 2.4	12, 13.2 ± 1.8
	11, 3.7 ± 1.0	11, 23.4 ± 2.2	14, 24.2 ± 2.2	7, 20.3 ± 2.1	8, .3 ± .3	12, 27.6 ± 2.4	11, 35.7 ± 2.6	13, 25.6 ± 2.3
	12, 12.4 ± 1.7	12, 29.9 ± 2.4	15, 22.1 ± 2.1	8, 7.7 ± 1.4	9, .3 ± .3	13, 7.8 ± 1.4	12, 23.2 ± 2.3	14, 25.3 ± 2.3
	13, 15.0 ± 1.8	13, 7.3 ± 1.4	16, 15.7 ± 1.9	9, 1.1 ± .6	10, 10.3 ± 1.6	14, .6 ± .4	13, 4.8 ± 1.2	15, 17.4 ± 2.0
	14, 6.8 ± 1.3	14, .8 ± .5	17, .8 ± .4	10, .6 ± .4	11, 4.0 ± 1.0			16, 6.5 ± 1.3
	15, 1.6 ± .6	15, .5 ± .4		11, 1.1 ± .6				17, .8 ± .5
	16, 1.0 ± .5			12, .3 ± .3				
				13, 2.0 ± .7				
				14, 1.1 ± .6				
				15, .3 ± .3				
	n 380	368	376	350	348	348	328	356
Mexican-Americans:								
	7, 39.6 ± 2.5	7, .3 ± .3	10, 1.1 ± .5	3.2,[a] 25.4 ± 2.3	5, 41.2 ± 2.6	9, 1.7 ± .7	8, .6 ± .4	9, .9 ± 0.5
	8, 17.9 ± 2.0	8, .3 ± .3	11, 36.2 ± 2.5	4, 10.1 ± 1.6	6, 28.3 ± 2.4	10, 53.6 ± 2.6	9, .3 ± .3	10, .9 ± .5
	10, .8 ± .4	4, 2.1 ± .7	12, 7.3 ± 1.4	5, 16.1 ± 1.9	6A,[a] .3 ± .3	11, 20.1 ± 2.1	10, 16.6 ± 2.2	11, 4.6 ± 1.2
	12, 37.0 ± 2.4	10, 29.1 ± 2.4	13, 9.2 ± 1.5	6, 19.9 ± 2.1	7, 4.7 ± 1.1	12, 21.0 ± 2.2	11, 43.4 ± 2.8	12, 12.2 ± 1.8
	13, 2.8 ± .8	11, 28.9 ± 2.3	14, 13.2 ± 1.8	7, 26.2 ± 2.3	8, .5 ± .4	13, 3.1 ± .9	12, 27.7 ± 2.5	13, 32.9 ± 2.6
	14, 1.8 ± .7	12, 31.6 ± 2.4	15, 23.8 ± 2.2	8, .8 ± .5	9, 1.1 ± .5	14, .6 ± .4	13, 7.8 ± 1.5	14, 17.7 ± 2.1
		13, 6.7 ± 1.3	16, 8.9 ± 1.5	13, .8 ± .5	10, 22.8 ± 2.2		14, 1.6 ± .7	15, 9.1 ± 1.6
		14, 1.1 ± .5	17, .3 ± .3	15, .5 ± .4	11, 1.1 ± .5			16, 19.8 ± 2.2
								17, 1.5 ± .7
								18, .3 ± .3
	n 386	374	370	366	364	358	318	328
Asians:								
	6, 2.6 ± 1.3	9, 4.9 ± 1.8	11, 29.5 ± 3.7	3.2,[a] 27.8 ± 3.4	5, 29.9 ± 3.7	9, 1.3 ± .9	8, .7 ± .7	9, 2.9 ± 1.4
	7, 56.4 ± 4.0	10, 27.8 ± 3.7	12, 3.2 ± 1.4	4, 8.7 ± 2.5	6, 17.5 ± 3.1	10, 67.5 ± 3.8	9, 1.5 ± 1.0	11, 8.8 ± 2.4
	8, 1.9 ± 1.1	11, 24.3 ± 3.6	13, 17.9 ± 3.1	5, 10.3 ± 2.7	7, 2.6 ± 1.3	11, 13.0 ± 2.7	10, 9.7 ± 2.6	12, 33.1 ± 4.0
	10, 1.3 ± .9	12, 33.3 ± 3.9	14, 21.8 ± 3.9	6, 50.8 ± 4.4	9, 1.9 ± 1.1	12, 13.0 ± 2.7	11, 47.0 ± 4.3	13, 41.2 ± 4.2
	11, .6 ± .6	13, 9.0 ± 2.4	15, 12.8 ± 2.7	7, .8 ± .8	10, 40.3 ± 4.0	13, 4.5 ± 1.7	12, 21.6 ± 3.6	14, 11.0 ± 2.7
	12, 35.9 ± 3.8	14, .7 ± .7	16, 14.7 ± 2.8	12, .8 ± .8	11, 6.5 ± 2.0	14, .6 ± .6	13, 18.7 ± 3.4	15, 1.5 ± 1.0
	13, .6 ± .6			16, .8 ± .8	12, 1.3 ± .9			
	14, .6 ± .6							
	n 156	144	156	126	154	154	134	136

NOTE.—Locus designations are the GenBank locus name, without the repeat designations. The allele designation is the no. of units of the core repeat. Allele frequency (%) ± standard error of the allele frequency is shown. *n* = No. of chromosomes sampled.

[a] Variant alleles that do not migrate with adjacent allele repeat reiterations. We have followed the recommendations on nomenclature standardization of STR loci, made by the DNA commission members at the meeting of the International Society of Forensic Haemogenetics in Venice, Italy, in October 1993, in naming these alleles, where the sequence of the variant is known (i.e., 3.2 is a fragment that migrates at the position of a fragment that is two bases longer than a fragment containing three repeats of the core sequence); unsequenced variants are designated with the adjacent smaller allele and a letter (i.e., 6A).

Appendix B

Table B1

Tests of Hardy-Weinberg Expectation

Locus and Test[a]	Population[b]				
	Caucasian	Black	Mexican-American	Asian	Pooled
HPRTB:					
χ^{2b}	.013[c] (.629)	.050[c] (.482)	.853	.808	.000[d]
L	.651	.755	.253	.329	.097
E	.357	.478	.170	.306	.038[c]
r	.058	.031	.048	.046	
FABP:					
χ^2	.577	.289	.080	.637	.005[c]
L	.905	.334	.610	.713	.000[d]
E	.835	.506	.310	.460	.000[d]
r	.000	.000	.019	.400	
CD4:					
χ^2	.935	.839	.509	.203	.001[d]
L	.225	.218	.567	.138	.000[d]
E	.407	.101	.516	.112	.000[d]
r	.000	.000	.006	.074	
CSF1PO:					
χ^2	.789	.347	.108	.269	.105
L	.833	.274	.803	.095	.078
E	.819	.376	.890	.080	.114
r	.000	.000	.000	.000	
TH01:					
χ^2	.266	.164	.068	.039[c] (.393)[b]	.013[c]
L	.085	.004[c] (.004)[b]	.457	.121	.002[d]
E	.062	.004[c] (.004)[b]	.314	.047[c] (.467)[b]	.001[d]
r	.030	.000	.025	.032	
PLA2A1:					
χ^2	.325	.486	.920	.000[d] (.486)[b]	.026[c]
L	.500	.432	.701	.078	.128
E	.629	.463	.819	.029[c] (.603)[b]	.108
r	.000	.000	.000	.086	
F13A01:					
χ^2	.037[c] (.257)[b]	.396	.643	.878	.000[d]
L	.599	.049[c] (.306)[b]	.461	.450	.021[c]
E	.536	.060	.670	.287	.011[c]
r	.016	.008	.000	.005	
CYAR04:					
χ^2	1.000	.006[d] (.070)[b]	.670	.224	.003[d]
L	.425	.009[d] (.135)[b]	.544	.971	.018[c]
E	.335	.006[d] (.366)[b]	.430	.979	.016[c]
r	.018	.016	.000	.000	
LIPOL:					
χ^2	.312	.350	.844	.379	.589
L	.467	.599	.999	.341	.374
E	.605	.594	.998	.603	.369
r	.000	.002	.000	.000	
RENA4:					
χ^2	.793	.226	.800	1.000	.396
L	.960	.281	.519	.062	.617
E	.963	.129	.391	.058	.211
r	.000	.007	.000	.019	
FESFPS:					
χ^2	.729	.373	.110	.007[d] (.162)[b]	.016[c]
L	.970	.461	.351	.146	.338
E	.991	.186	.442	.041[c] (.455)[b]	.111
r	.000	.029	.009	.039	

Table B1 (continued)

Locus and Test[a]	Population[b]				
	Caucasian	Black	Mexican-American	Asian	Pooled
D6S366:					
χ^2	.099	.542	.678	.106	.003[c]
L	.454	.052	.213	.163	.302
E	.307	.036[c] (.411)[b]	.281	.063	.151
r	.008	.009	.000	.031	
ARA:					
χ^2	.577	.013[c] (.757)[b]	.441	.084	.000[d]
L	.248	.065	.426	.244	.013[c]
E	.224	.015[d] (.383)[b]	.266	.143	.004[c]
r	.006	.033	.000	.060	

[a] The empirical levels of significance (based on 5,000 replications of allele shuffling), in the three test procedures for detecting derivations of the observed genotype frequencies from those expected under Hardy-Weinberg expectations, were calculated. The tests were conducted for genotypes of females only for the X-linked loci. χ^2 analysis is of observed and expected no. of heterozygotes; probability values reflect 5,000 permutations. Probabilities computed from the L test ($-2 \ln L_0/L_1$) were determined with 5,000 permutations.

[b] Nos. in parentheses are revised levels of significance, incorporating the estimated rs in shuffling of alleles across individuals, shown only when the results were significant.

[c] Significant deviation from Hardy-Weinberg proportions, $P < .05$.

[d] Significant deviation from Hardy-Weinberg proportions, $P < .01$.

References

Ahn YI, Kamboh MT, Ferrell RE (1992) Two new alleles in the tetranucleotide repeat polymorphism at the lipoprotein lipase (LPL) locus. Hum Genet 90:184

Bell GI, Selby MJ, Rutter WJ (1982) The highly polymorphic region near the insulin gene is composed of simple tandemly repeating sequences. Nature 295:31–35

Boylan KB, Ayres TM, Popko B, Takahashi N, Hood LE, Prusiner SB (1990) Repetitive DNA $(TGGA)_n$ 5′ to the human myelin basic protein gene: a new form of oligonucleotide repetitive sequence showing length polymorphism. Genomics 6:16–22

Brown AHD, Feldman MW, Nevo E (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics 96: 523–536

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR (1993) Incidence and origin of "null" alleles in the (AC)n microsatellite markers. Am J Hum Genet 52:922–927

Caskey CT, Hammond HA (1992) Forensic use of short tandem repeats *via* PCR. In: Advances in forensic haemogenetics. Springer, Berlin, pp 18–25

Chakraborty R (1984) Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. Genetics 108:719–731

——— (1993*a*) A class of population genetic questions formulated as the generalized occupancy problem. Genetics 134: 953–958

——— (1993*b*) Generalized occupancy problem and its application in population genetics. In: Sing CF, Hanis CL (eds) Genetics of cellular, individual, family, and population variability. Oxford University Press, New York, pp 179–192

Chakraborty R, Fornage M, Guegue R, Boerwinkle E (1991) Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In: Burke T, Dolf G, Jeffreys AJ, Wolff R (eds) DNA fingerprinting: approaches and applications. Birkhäuser, Basel, Boston, Berlin, pp 127–143

Chakraborty R, Kidd K (1991) The utility of DNA typing in forensic work. Science 254:1735–1739

Chakraborty R, Neel JV (1989) Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. Proc Natl Acad Sci USA 86:9407–9411

Chakraborty R, Zhong Y, Jin L, Budowle B (1994) Nondetectability of restriction fragments and independence of DNA-fragment sizes within and between loci in RFLP typing of DNA. Am J Hum Genet 55 (in press)

Chakravarti A, Li CC (1983) The effect of linkage on paternity calculations. In: Walker RH (ed) Inclusion probabilities in parentage testing. American Association of Blood Banks, Arlington, VA, pp 411–420

Deka R, Chakraborty R, Ferrell RE (1991) A population genetic study of six VNTR loci in three ethnically defined populations. Genomics 11:83–92

Edwards A, Civitello A, Hammond HA, Caskey CT (1991*a*) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet 49:746–756

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem

repeat loci in four human population groups. Genomics 12: 241–253

Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, et al (1990) Automated DNA sequencing of the human HPRT locus. Genomics 6:593–608

Edwards MC, Clemens PR, Tristan M, Pizzuti A, Gibbs RA (1991*b*) Pentanucleotide repeat length polymorphism at the human CD4 locus. Nucleic Acids Res 19:4791

Gart JJ, Nam J (1984) A score test for the possible presence of recessive alleles in generalized ABO-like genetic systems. Biometrics 40:887–894

Guo S-W, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hearne CM, Todd JA (1991) Tetranucleotide repeat polymorphism at the HPRT locus. Nucleic Acids Res 19:5450

Huang TH-M, Hejtmancik JF, Edwards A, Pettigrew AL, Herrera CA, Hammond HA, Caskey CT, et al (1991) Linkage of the gene for an X-linked mental retardation disorder to a hypervariable $(AGAT)_n$ repeat motif within the human hypoxanthine phosphoribosyltransferase (HPRT) locus (Xq26). Am J Hum Genet 49:1312–1319

Jeffreys AJ, Royle NJ, Wilson V, Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332:278–281

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73

Koorey DJ, Bishop GA, McCaughan GW (1993) Allele non-amplification: a source of confusion in linkage studies employing microsatellite polymorphisms. Hum Mol Genet 2:289–291

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Milkman R, Beatty LD (1970) Large-scale electrophoresis studies of allelic variation in Mytilus Edulis. Biol Bull 139:430

Morton NE, Collins A, Balazs I (1993) Kinship bioassay on hypervariable loci in Blacks and Caucasians. Proc Natl Acad Sci USA 90:1892–1896

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Panzer SW, Hammond HA, Stephens L, Chai A, Caskey CT (1993) Trinucleotide repeat polymorphism at the D6S366 locus. Hum Mol Genet 2:1511

Parentage Testing Committee, American Association of Blood Banks (1990) Standards for parentage testing laboratories. American Association of Blood Banks, Arlington, VA

Patel PI, Nussbaum RL, Gramson PE, Ledbetter DH, Caskey CT, Chinault AC (1984) Organization of the HPRT gene and related sequences in the human genome. Somat Cell Mol Genet 10:483–493

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1990) Trinucleotide repeat at the human pancreatic phospholipase A-2 gene (PLA2). Nucleic Acids Res 18:7468

——— (1991*a*) Tetranucleotide repeat polymorphism at the human c-fes/fps proto-oncogene (FES). Nucleic Acids Res 19: 4018

——— (1991*b*) Tetranucleotide repeat polymorphism at the human coagulation factor XIII A subunit gene (F13A1). Nucleic Acids Res 19:4306

——— (1991*c*) Trinucleotide repeat polymorphism at the human fatty acid binding protein gene (FABP2). Nucleic Acids Res 18: 7198

Polymeropoulos MH, Xiao H, Rath DS, Merril CR (1991*d*) Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (TH). Nucleic Acids Res 19:3753

——— (1991*e*) Tetranucleotide repeat polymorphism at the human aromatase cytochrome P-450 gene (CYP19). Nucleic Acids Res 19:195

Puers C, Hammond HA, Jin L, Caskey CT, Schumm JW (1993) Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01$[AATG]_n$ and reassignment of alleles in population analysis by using a locus-specific allelic ladder. Am J Hum Genet 53:953–958

Risch NJ, Devlin B (1992) On the probability of matching DNA fingerprints. Science 255:717–720

Roewer L, Epplen JT (1992) Rapid and sensitive typing of forensic stains by PCR amplifications of polymorphic simple repeat sequences in case work. Forensic Sci Int 53:163–171

Sensabaugh GF (1982) Biochemical markers of individuality. In: Saferstein R (ed) Forensic science handbook. Prentice-Hall, Englewood Cliffs, NJ, pp 338–415

Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993

Sleddens HF, Oostra BA, Brinkmann AO, Trapman J (1992) Trinucleotide repeat polymorphism in the androgen receptor gene (AR). Nucleic Acids Res 20:1427

Sokal RR, Rohlf JF (1969) Biometry, 2d ed. WH Freeman, New York

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17: 6463–6471

Technical Working Group on DNA Analysis Methods, California Association of Criminalists Ad Hoc Committee on DNA Quality Assurance (1991) Guidelines for a quality assurance program for DNA analysis. Crime Lab Dig 18:44–75

Weber JL (1990) Informativeness of human $(dC\text{-}dA)_n$-$(dG\text{-}dT)_n$ polymorphisms. Genomics 7:524–530

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weir BS (1991) Genetic data analysis. Sinauer, Sunderland, MA

Yasuda N (1968) Estimation of the inbreeding coefficient from phenotype frequencies by a method of maximum likelihood scoring. Biometrics 24:915–935

Zuliani G, Hobbs HH (1990) Tetranucleotide repeat polymorphism in the LPL gene. Nucleic Acids Res 18:4958